Pergamon

PII: S0098-3004(97)00049-6

Computers & Geosciences Vol. 23, No. 7, pp. 763–770, 1997

Printed in Great Britain
0098-3004/97 $17.00 + 0.00

A SAS-IML PROGRAM FOR IMPLEMENTING TWO-PHASE REGRESSION ANALYSIS OF GEOPHYSICAL TIME SERIES DATA

GERRI M. DUNNIGAN,[1] JOHN L. HAMMEN,[2] and T. ROBERT HARRIS[1]

Department of Mathematics, Box 8376, University of North Dakota, Grand Forks, North Dakota 58202, U.S.A., and [2]Department of Geography, Box 9020, University of North Dakota, Grand Forks, North Dakota 58202, U.S.A.
(*e-mail*: gerrid@sage.und.nodak.edu)

(Received 29 May 1996; revised 17 February 1997)

Abstract—Two-phase regression analysis has been shown to have utility in geophysical time series analysis. Based on linear regression, the technique operates by locating a change point, if one exists, where a significant change in slope occurs. The timing of the break can then be associated with natural and anthropogenic variables which are thought to impact the behavior of the dependent variable. The technique is not widely available in commercial statistical packages. A SAS Interactive Matrix Language program is presented here to implement the technique. © 1997 Elsevier Science Ltd

Key Words: Linear regression, Two-phase regression, Change point, SAS-IML.

INTRODUCTION

A variety of techniques exists for time series analysis of geophysical data, with differing degrees of sophistication and utility. Linear regression, where a geophysical variable (e.g. temperature) is regressed over time, is one of the most basic and widely used deterministic models. Questions of appropriateness and applicability aside (Woodward and Gray, 1993; Garbrecht and Fernandez, 1994), there are additional significant considerations in using linear regression either to predict future trends or to describe past events mathematically. Specifically, linear regression may "mask" other effects in the data (e.g. autocorrelation) or may be significantly influenced by abrupt data "breaks". Kite (1993), for example, in evaluating secular trends in lake levels at Lake Victoria, Jinja, noted that in initial tests, 30% of the variation was a result of linear trend. However, when a period of sudden rise (four years in duration) was removed from the data set, variance due to linear trend became negligible. These data "breaks" are of particular interest where attempts are made to correlate changes in data patterns with the timing of other environmental factors or anthropogenic activity.

Two models which provide valuable insight into the existence, timing, and significance of data "breaks" are two-phase and piecewise regression. Conceptually, the techniques are extensions of linear regression wherein change points are identified as locations where the regression line changes. A linear regression is then calculated for each subset of data (i.e. from the beginning of the data set to the change point and from the change point to the end of the data set). Two-phase regression mandates that the two best-fit lines be joined at the change point; there is always, then, a difference in slope between the two "new" data sets. Piecewise regression does not require that trend lines be joined at the change point, thus allowing any change in slope and intercept (Fig. 1).

Both techniques are enjoying increasing use, particularly in climate change scenarios. Hanson, Maul, and Karl (1989), for example, employed two-phase regression to test for constancy of mean in precipitation and temperature data averaged for the contiguous U.S. and the northern plains region of the U.S. Skaggs and Baker (1989) analyzed a long-term temperature record for eastern Minnesota using two-phase regression. Cooter and Leduc (1995) utilized piecewise regression to detect discontinuities in frost dates in the north-eastern U.S.

The techniques provide a valuable, "second-step" analytical procedure, easily adopted and understood by those familiar with basic linear regression. In addition, whereas the concepts surrounding two-phase regression are described in the literature, it may not be clear to applied scientists how to implement the computations needed for estimates and significance tests. Therefore we offer a two-phase regression program readily implemented using the Interactive

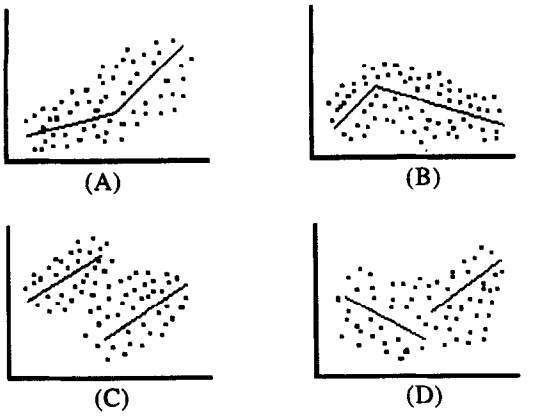

Figure 1. Two-phase and piecewise regression. (A) and (B) illustrate required connection between trendlines in two-phase regression, whereas (C) and (D) show discontinuities allowed in piecewise regression.

Matrix Language (IML) available in both mainframe- and PC-based SAS.

TWO-PHASE REGRESSION

The two-phase model can be written:

$$y_i = f(x_i; \alpha_0, \beta_0, \alpha_1, \beta_1, c) + e_i \ (i = 1, 2, \ldots, m) \quad (1)$$

where

$$f(x_i; \alpha_0, \beta_0, \alpha_1, \beta_1, c) = \begin{cases} \alpha_0 + \beta_0 x_i & \text{if } x_i < c \\ \alpha_1 + \beta_1 x_i & \text{if } x_i \geq c, \end{cases} \quad (2)$$

e_1, e_2, ..., e_m are independent, normally distributed random errors with mean 0 and common variance σ^2; and

$$\alpha_0 + \beta_0 c = \alpha_1 + \beta_1 c. \quad (3)$$

In the applications discussed here, x_1, x_2, ..., x_m are times and are usually equally spaced. Without loss of generality, $x_i = i$ (with the possible exception of a few missing values). Thus the values y_1, y_2, ..., y_m are a time series with the trend given by Equation (1). The independence, homoscedasticity, and normality of the errors allow the use of an extension of ordinary least squares (OLS) regression methods. The user of these methods will need to evaluate the appropriateness of these assumptions for the application. In particular, the assumption of uncorrelated errors in time series has often been made (Solow, 1987) but has also been deemed inappropriate in many situations (Box, Jenkins, and Reinsel, 1994).

The unknown parameters are the regression line parameters α_0, β_0, α_1, and β_1; the change point parameter c; and the error variance σ^2. Equation (3) implements the restriction of the two-phase model, that the lines intersect at the point $x = c$. Thus given c, only three of the four regression line parameters are functionally independent. After algebraic manipulation, the model can be written:

$$y_i = a + b_0 x_i + b w_i + e_i \quad (4)$$

where

$$\alpha_0 = a, \ \beta_0 = b_0, \ \beta_1 = b_0 + b \quad (5)$$

and

$$w_i = \begin{cases} x_i - c & \text{if } x_i \geq c \\ 0 & \text{if } x_i < c. \end{cases} \quad (6)$$

In this form, the parameters are regression line parameters a, b_0, and b; change point parameter c; and error variance σ^2. The model is nonlinear because of the dependence on c.

In either form, the model may be estimated by fixing c, obtaining the (usual) OLS estimates of the regression line parameters given c, and computing the residual sum of squares, which depends on c. The value of c which minimizes the residual sum of squares can be determined by a numerical method such as grid search. The OLS estimates are then:

$\hat{c}$: the value of c that minimizes the residual sum of squares; (7)

$\hat{a}$, $\hat{b}_0$, $\hat{b}$: the OLS estimates of a, b_0, and b with c fixed at $\hat{c}$; and (8)

$\hat{\alpha}_0$, $\hat{\beta}_0$, $\hat{\alpha}_1$, $\hat{\beta}_1$: obtained by replacing a, b_0, and b by their estimates in Equation(5). (9)

The two-phase model may be compared with the simple linear regression model

$$y_i = \alpha + \beta x_i + e_i \quad (10)$$

by the generalized likelihood ratio hypothesis test of the simple linear regression model as the null hypothesis, versus the two-phase model as the alternative hypothesis. (Thus the alternative hypothesis states that $b \neq 0$ and $\min(x_1, x_2, \ldots, x_m) < c < \max(x_1, x_2, \ldots, x_m)$). This procedure rejects the null hypothesis at significance level α if $U > F_{3,m-4}(\alpha)$, where

$$U = \frac{(S_0 - S)}{3} \Big/ \frac{S}{(m-4)};$$

S_0 is the residual sum of squares of the simple linear regression model; S is the residual sum of squares of the two-phase model; and $F_{3,m-4}(\alpha)$ is the upper α critical value of the F distribution with 3 and $m-4$ degrees of freedom. The degrees of freedom may seem surprising in view of the number of parameters in the two models, but they have been verified empirically by Hinkley (1971). Apparently they are affected by the nonlinearity of the two-phase model.

IML PROGRAM DESCRIPTION AND IMPLEMENTATION

The following program will fit a simple linear model and a two-phase regression model to a set of observations and then test whether the two-phase model is significantly better than the simple linear regression model. The program is designed for data in which the independent variable is integer valued, as is the situation for many geophysical applications where observations or measurements are made yearly. Slight modifications would make it suitable for data in which the independent variable assumes non-integer values. The program presented next is applied to the Baker, Watson, and Skaggs (1985) long-term temperature record data for eastern Minnesota. Temperature is measured in degrees Celsius.

It is assumed that the data are stored in a file with each pair of observations (year and temperature in this application) entered on a separate line with at least one space separating the two observations. If the data file is formatted differently, then it needs to be altered or appropriate changes to the data step will be required. Here the file `Baker` contains the year and temperature values for eastern Minnesota for the years 1820 to 1982.

```
data set1;
  infile Baker;
  input Year Temp;
```

The independent variable, `Year` in this application, is scaled so that it has a minimum value of one. In this application the first recorded temperature corresponds to the year 1820. So by subtracting 1819 from all values of `Year` in the data set, the independent variable will have values ranging from 1 to 163. Scaling the independent variable is conducive to more general program statements and less demanding computations. `ScalYr` is the name given to the scaled `Year` variable. The following code creates a data set termed `MinnTemp` containing the variables `ScalYr` and `Temp` and prints the data set.

```
data MinnTemp; set set1;
  ScalYr=Year - 1819;
  proc print data=MinnTemp;
```

This program uses the SAS IML (Interactive Matrix Language) (1989) programming language to fit the simple linear and two-phase regression models. The following code invokes IML and loads the `MinnTemp` data into the matrix D. Matrix D is then printed for reference.

```
proc iml;
  use MinnTemp;
  read all var {ScalYr Temp} into D;
print D;
```

A simple grid search is used in fitting the two-phase regression model. The range of values for the independent variable, in this application 1 to 163, is partitioned into equally spaced subintervals. The space between successive points in the partition will be $1/2^p$ for some integer p chosen by the user. For example, if $p = 2$, the partition would consist of the points 1, 1.25, 1.5, 1.75, 2, 2.25, ..., 162.5, 162.75, 163. The program considers in turn each interior point of the partition (excluding the endpoints 1 and 163) as the change point, c, and computes the corresponding residual sum of squares. The program then determines which point yields the smallest residual sum of squares and takes this point as the estimate of the change point. (This estimate was denoted $\hat{c}$ in the preceding discussion of two-phase regression but will be output as c by the program. The choice of p is left to the user of the program). N is the largest value of `ScalYr` and M is the total number of observations in the data set. Note that N and M will be the same if there are no missing values in the data set as is the situation for this application where $N = M = 163$. G is the number of interior points in the partition.

```
N=163;
M=163;
p=0;
G=(N-1)*(2**p)-1;
```

The following code defines the matrices to be used in fitting the two-phase regression models (using standard regression notation) as the change point c assumes values in the grid. The first three lines set the number of rows and columns in each of the matrices X, Y, and SSE, and the remaining lines load data into X and Y from the data matrix D.

```
X=j(M,3,0);          /* Design matrix X has M rows and 3 columns */
Y=j(M,1,0);          /* Dependent variable matrix Y has M rows and 1 column */
SSE=j(G+1,5,0);        /* Matrix SSE has G+1 rows and 5 columns */
do i=1 to M;
  X[i,1]=1;            /* Intercept */
  X[i,2]=D[i,1];       /* ScalYr */
  Y[i,1]=D[i,2];       /* Temp */
end;
```

The following code computes the parameter estimates and residual sum of squares obtained when fitting a two-phase regression model to the data using each of the grid values for c, and places them in the matrix *SSE*. The first column of *SSE* will contain the values of c, the second column will contain the corresponding residual sum of squares, and columns three through five will contain the corresponding parameter estimates $\hat{a}$, $\hat{b}_0$, and $\hat{b}$, respectively, from Equation (8).

```
do k=1 to G;
  c=(2**p+k)/2**p;
  do i=1 to M;              /* w_i in (4) */
    if D[i,1]<=c then X[i,3]=0;
    else X[i,3]=D[i,1]-c;
  end;
  SSE[k,1]=c;               /* c values go into column 1 */
  b=inv(X'*X)*X'*Y;
  SSE[k,2]=Y'*Y-b'*X'*Y;    /* Residual sums of squares go into column 2 */
  SSE[k,3]=b[1,1];          /* â */
  SSE[k,4]=b[2,1];          /* b̂0 */
  SSE[k,5]=b[3,1];          /* b̂ */
end;
```

Matrix X is then redefined and a simple linear regression model is fitted to the data. The corresponding residual sum of squares becomes the last entry in column two of the matrix *SSE*. Matrix *SSE* is then printed for reference and a plot of the residual sum of squares is constructed.

```
X=j(M,2,0);   /* Matrix X has M rows and 2 columns in fitting simple model */
do i=1 to M;
  X[i,1]=1;          /* Intercept */
  X[i,2]=D[i,1];     /* ScalYr */
end;
SSE[G+1,1]=1;                  /* "c=1" denotes the simple linear model */
b=inv(X'*X)*X'*Y;
SSE[G+1,2]=Y'*Y - b'*X'*Y;     /* Residual sum of squares for simple model */
Alpha=b[1,1];                 /* α̂ estimates α in (10) */
Beta=b[2,1];                  /* β̂ estimates β in (10) */
print SSE;
```

The following code constructs a plot of the residual sum of squares. Because `pgraf` (which is used to plot the residual sum of squares against c) operates on matrices with two columns, we first define the matrix S as a matrix containing the first two columns of *SSE*.

```
S=j(G+1,2);  /*Matrix S will contain the first two columns of SSE */
do i=1 to G+1;
  S[i,1]=SSE[i,1];
  S[i,2]=SSE[i,2];
end;
call pgraf(S,'$','c','SSE','Residual Sum of Squares');
```

A search for the minimum residual sum of squares value is then conducted. The minimum value is assigned to the variable `MinSSE`; c assumes the value of the grid point that corresponds to this minimum value; and a, b_0, and b will assume the values of the parameter estimates associated with this value of c.

```
c=1;
MinSSE=SSE[G+1,2];
SLSSE=SSE[G+1,2];   /* Residual sum of squares for the simple linear model */
do k=1 to G;
  if SSE[k,2]<=MinSSE then do;
    c=SSE[k,1];
    MinSSE=SSE[k,2];
    a=SSE[k,3];
    b0=SSE[k,4];
    b=SSE[k,5];
  end;
end;
```

The parameter estimates and residual sum of squares for the two-phase and simple linear regression models are printed.

```
print Alpha;
print Beta;
print SLSSE;
print a;
print b0;
print b;
print c;
print MinSSE;
```

A hypothesis test of the two-phase regression model versus the simple linear regression model is then performed. The test statistic, `Ftest`, and its corresponding p-value, `Pval`, are then printed.

```
DDF=M-4;
Ftest=((SSE[G+1,2] - MinSSE)/3)/(MinSSE/DDF);
print Ftest;
Pval=1 - probf(Ftest,3,DDF);
print Pval;
quit;
```

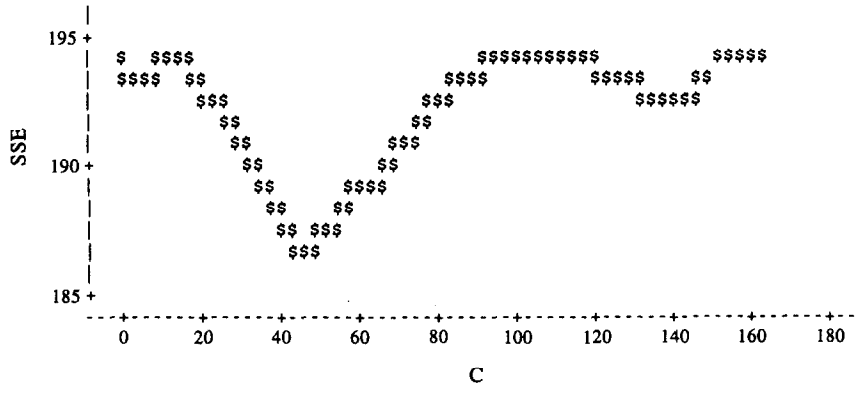

Figure 2. Plot of residual sum of squares with respect to change point value c.

PROGRAM OUTPUT

In this application to the Baker, Watson, and Skaggs (1985) data, the following output was produced. Figure 2 is a plot of the residual sum of squares plotted against the values of c in the grid. This is a rough plot, but one can get a good idea of the behavior of the residual sum of squares for the different choices of change point. In this example, one can clearly see the approximate location of the minimum. In some applications, there may be more than one significant "valley" in the graph. This would indicate the presence of a local minimum or local minima which may warrant further investigation. These local minima may be identified by looking at the residual sum of squares plot, even though the program (excluding the plot) only identifies the absolute minimum sum of squares and the corresponding change point c. The selection of p warrants further discussion. In this application, it was determined that $p = 0$ (which yields a partition containing just the integers from 1 to 163) gave the same results as did larger values of p. Further applications involving independent variables assuming only integer values yielded the same result. However, in an application involving an independent variable whose values were not restricted to integers, $p = 8$ was required before no change was observed in the output value of the change point c. One possible strategy, then, is to run the program consecutively using larger values of p until no change is observed in the output value of c. One may also choose values of p to achieve a predetermined level of precision.

Table 1. Parameter estimates and hypothesis test results.

	Simple Model	Two-Phase Model
SSE	(*SLSSE*) 194.400	(*MinSSE*) 186.955
Intercept(s)	$\hat{\alpha}$(*Alpha*) = 5.540	$\hat{\alpha}_0(A) = 6.136$, $\hat{\alpha}_1 = 4.961$
Slope(s)	$\hat{\beta}$(*Beta*) = 0.007	$\hat{\beta}_0(B0) = -0.012$, $\hat{\beta}_1 = 0.013$
Change Point		$\hat{c}(C) = 47$
Test Statistic		U (*FTest*) = 2.111
p-value		(*PVal*) = 0.101

The parameter estimates and hypothesis test results are summarized in Table 1. The notation for these estimates and test results used in the program are given in parentheses. The values of the parameter estimates not output by the program are computed using formulas given in the theory section and the parameter estimates output by the program. The estimate of the change point is the same and the parameter estimates are nearly the same as those obtained by Skaggs and Baker (1989). The differences observed in the parameter estimates, the sum of squares, the test statistic, and the p-value are attributable to five additional data points that were used by Skaggs and Baker (1989) that were not yet recorded by Baker, Watson, and Skaggs (1985). In this application, the test indicates that the two-phase model is not significantly better than the linear model at significance level $\alpha = 0.10$ or lower.

SUMMARY

Two-phase regression analysis is a sophisticated yet, as presented here, an easily implemented technique of considerable utility in time series analysis. It is especially well suited as a preliminary analytical step in studies attempting to discern changes in geophysical data, wherein the time of change is of interest. The technique has also been used to remove serial correlation effects from a geophysical dataset (Skaggs, Baker, and Ruschy, 1995). The code is publicly available by anonymous FTP from IAMG.ORG.

REFERENCES

Baker, D. L., Watson, B. F. and Skaggs, R. H. (1985) The Minnesota longterm temperature record. *Climatic Change* 7, 225–236.

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (1994) *Time Series Analysis: Forecasting and Control.* Prentice-Hall, Englewood Cliffs, New Jersey, 598 pp.

Cooter, E. J. and Leduc, S. K. (1995) Recent frost date trends in the north-eastern USA. *International Journal of Climatology* **15**, 65–75.

Garbrecht, J. and Fernandez, G. P. (1994) Visualization of trends and fluctuations in climatic records. *Water Resources Bulletin* **30**(2), 297–306.

Hanson, K., Maul, G. A. and Karl, T. R. (1989) Are atmospheric "greenhouse" effects apparent in the climatic record of the contiguous U.S. (1895–1987)?. *Geophysical Research Letters* **16**(1), 49–52.

Hinkley, D. V. (1971) Inference in two-phase regression. *Journal of the American Statistical Association* **66**(336), 736–743.

Kite, G. (1993) Analyzing hydrometeorological time series for evidence of climatic change. *Nordic Hydrology* **24**, 135–150.

SAS Institute Inc., (1989) *SAS/IML® Software: Usage and Reference, Version* 6. 1st edn. SAS Institute Inc., Cary, NC, 501 pp.

Skaggs, R. H. and Baker, D. L. (1989) Temperature change in eastern Minnesota. *Journal of Climate* **2**, 629–630.

Skaggs, R. H., Baker, D. L. and Ruschy, D. L. (1995) Interannual variability characteristics of the eastern Minnesota (USA) temperature record: implications for climate change studies. *Climate Research* **5**, 223–227.

Solow, A. R. (1987) Testing for climate change: an application of the two-phase regression model. *Journal of Climate and Applied Meteorology* **26**, 1401–1405.

Woodward, W. A. and Gray, H. L. (1993) Global warming and the problem of testing for trend in time series data. *Journal of Climate* **6**, 953–962.